

Two-Sided Generalized Topp and Leone (TS-GTL) distributions

Donatella Vicari, Department of Statistics, Probability and Applied Statistics, University of Rome "La Sapienza", Rome, Italy. E-mail: donatella.vicari@uniroma1.it

Corresponding Author:

Johan Rene van Dorp, Department of Engineering Management and Systems Engineering, School of Engineering and Applied Science, The George Washington University, 1776 G Street, N.W. , Washington D.C., 20052, USA. E-mail: dorpjr@gwu.edu. Phone: 202-994-6638. Fax: 202-994-0245.

Samuel Kotz, , Department of Engineering Management and Systems Engineering, School of Engineering and Applied Science, The George Washington University, 1776 G Street, N.W. , Washington D.C., 20052, USA. E-mail: kotz@gwu.edu

Short running title: TS-GTL distributions

Date of Manuscript Submission : October 17, 2007

Revised : May 7, 2008

Two-Sided Generalized Topp and Leone (TS-GTL) distributions

Donatella Vicari¹, Johan Rene van Dorp² and Samuel Kotz³

Abstract — Over 50 years ago in a 1955 issue of JASA a paper on a bounded continuous distribution by Topp and Leone (1955) has appeared. (The subject was subsequently dormant for over 40 years but recently the family was resurrected.) Here we shall investigate the so-called Two-Sided Generalized Topp and Leone (TS-GTL) distributions. This family of distributions is constructed by extending the Generalized Two-Sided Power (GTSP) family to a new two-sided framework of distributions, where the first (second) branch arises from the distribution of the largest (smallest) order statistic. The TS-GTL distribution is generated from this framework by sampling from a slope (reflected slope) distribution for the first (second) branch. The resulting five-parameter TS-GTL family of distributions turns out to be flexible, encompassing the uniform, triangular, GTSP and two-sided slope distributions into a single family. In addition, the pdf's may be having bimodal shapes or admitting shapes with a jump discontinuity at the "threshold" parameter. We discuss some properties of the TS-GTL family and describe a maximum likelihood estimation procedure. A numerical example of the MLE procedure is provided by means of a bimodal Galaxy M87 data set concerning V-I color indices of 80 globular clusters. A comparison with a Gaussian mixture fit is presented.

Mathematics Subject Classification (2000). Primary 60E05; Secondary 62H12

Keywords: Bimodal distribution, maximum likelihood estimation, order statistics

¹ Department of Statistics, Probability and Applied Statistics, University of Rome "La Sapienza", Rome, Italy

² Corresponding Author, Department of Engineering Management and Systems Engineering, The George Washington University, Washington D.C., USA

³ Department of Engineering Management and Systems Engineering, The George Washington University, Washington D.C., USA

1. Introduction

In 1919, E. Pairman and K. Pearson investigated continuous distributions on a limited range, including a particular estimation of their moments. Nevertheless, even in the late nineties of the 20th century relatively few probabilistic models of this kind were presented in the literature. Amongst them, the uniform, triangular and beta distributions are the most widely explored and applied, interspersed by some "curious" distributions (posed occasionally as problems or exercises). Other bounded continuous distributions were based on mathematical transformations of the normal distribution (possessing an unbounded domain) - the most wide-spread amongst them being the Johnson S_B family of transformations introduced in 1949. The multitude of existing unbounded continuous distributions developed during the 20th century contrasts with the relative scarcity of the bounded distributions.

This state of affairs motivated Van Dorp and Kotz (2003a) to study other constructs for families of distributions with bounded support. Initially, Van Dorp and Kotz (2003a) deal with the following family of two sided distributions

$$Pr(X \leq x|\theta, F(\cdot)) = \begin{cases} \theta F(\frac{x}{\theta}), & \text{for } 0 < x < \theta, \\ 1 - (1 - \theta)F(\frac{1-x}{1-\theta}), & \text{for } \theta \leq x < 1, \end{cases} \quad (1)$$

where $F(\cdot)$ is a *generating* cumulative distribution function (cdf) with bounded support $[0, 1]$ with the following condition $f(1) = 1$ on its pdf to ensure continuity of the pdf of (1) at the threshold parameter θ . Next, they also present in Kotz and Van Dorp (2004a) the following family of distributions:

$$Pr(X \leq x|\gamma, \Theta) = \begin{cases} p(\gamma, \Theta) \left(\frac{x}{\theta}\right)^m, & \text{for } 0 < x < \theta, \\ 1 - \{1 - p(\gamma, \Theta)\} \left(1 - \frac{x-\theta}{1-\theta}\right)^n, & \text{for } \theta \leq x < 1, \end{cases} \quad (2)$$

where $0 \leq \theta \leq 1$, $m > 0$, $n > 0$, $\gamma > 0$, the vector $\Theta = (\theta, m, n)$,

$$p(\gamma, \Theta) = Pr(X \leq \theta|\gamma, \Theta) = \frac{\gamma\theta n}{(1 - \theta)m + \gamma\theta n},$$

TS-GTL distributions

and parameter values $\gamma \neq 1$ allow for discontinuity of its pdf at the threshold parameter θ . The parameters m and n are evidently the *power parameters*. Note that, the families (1) and (2) are quite distinct in the sense that the left and right branch in (1) utilize the same generating cdf $F(\cdot)$ but does not allow for different power parameters m and n appearing in the left and right branch of (2). Other related contributions appearing prior to this paper have been summarized in the monograph by Kotz and Van Dorp (2004b) and the paper at hand should be viewed as a continuation of that work.

Specifically, we shall construct a novel single framework of families of distributions from (1) and (2) that combines previously separate families presented in Kotz and Van Dorp (2004b). Amongst these families are 1) Two-Sided Power Distributions (see also, Van Dorp and Kotz, 2002) 2) Two-Sided Slope Distributions and 3) Topp and Leone Distributions. All these three families share the uniform distribution and the right and left triangular distributions. Our starting point here is the three-parameter Generalized Two-Sided Power (GTSP) family of distributions (Kotz and Van Dorp, 2004a) obtained by substituting $\gamma = 1$ in (2). Certain generalizations below yield Two-Sided Generalized Topp and Leone (TS-GTL) distributions. This family constitutes a larger class of distribution families in the sense that it combines the previously available constructs in a single family, in particular allowing for bimodal forms of its pdf not covered in Kotz and van Dorp (2004b).

The remainder of this paper is organized as follows. In Section 2, we present our novel framework for two-sided distributions derived from (1) and (2). In Section 3, we utilize this framework to construct the TS-GTL distributions and present a variety of different shapes which can be taken by the pdf's of this new family. In Section 4, the moments are discussed. A maximum likelihood procedure for the TS-GTL distribution is derived in Section 5. Finally, we provide a numerical example of the MLE procedure applied to a recently acquired bimodal data set and compare the result with a Gaussian mixture fit.

2. A novel framework for two-sided distributions

The function x/θ [function $1 - (x - \theta)/(1 - \theta)$] in the first [second] branch of (2) is the cdf [reliability function] of a uniform random variable on $(0, \theta)$ [on $(\theta, 1)$]. This results in the following generalization of (2) using continuous cdf's $G(\cdot)$ and $H(\cdot)$ with the support $[0, 1]$ (while simultaneously substituting $\gamma = 1$ in (2)):

$$Pr(Y \leq y) = \begin{cases} p(\Theta) \left\{ G\left(\frac{y}{\theta}\right) \right\}^m, & \text{for } 0 < y < \theta, \\ 1 - \{1 - p(\Theta)\} \left\{ 1 - H\left(\frac{y-\theta}{1-\theta}\right) \right\}^n, & \text{for } \theta \leq y < 1, \end{cases} \quad (3)$$

The pdf corresponding to the cdf (3) is thus

$$f_Y(y|\Theta) = \frac{mn}{(1-\theta)m + \theta n} \begin{cases} g\left(\frac{y}{\theta}\right) \left\{ G\left(\frac{y}{\theta}\right) \right\}^{m-1}, & \text{for } 0 < y < \theta, \\ h\left(\frac{y-\theta}{1-\theta}\right) \left\{ 1 - H\left(\frac{y-\theta}{1-\theta}\right) \right\}^{n-1}, & \text{for } \theta \leq y < 1, \\ 0, & \text{elsewhere,} \end{cases} \quad (4)$$

(where $g(\cdot)$ and $h(\cdot)$ are the pdf's of the cdf's $G(\cdot)$ and $H(\cdot)$, respectively). While the cdf (3) is continuous at θ with value

$$p(\Theta) = \theta n / \{(1 - \theta)m + \theta n\}, \quad (5)$$

it follows from (4) that

$$f_Y(\theta^+|\Theta) - f_Y(\theta^-|\Theta) = h(0) - g(1), \quad (6)$$

and hence it is not necessarily continuous at θ . Moreover, $f_Y(\theta^-|\Theta) > 0$ [$f_Y(\theta^+|\Theta) > 0$] if and only if $g(1) > 0$ [$h(0) > 0$]. When the cdf's $G(\cdot)$ and $H(\cdot)$ coincide, the density $f_Y(y|\Theta)$ becomes continuous at θ and strictly positive at θ if and only if both $g(1) = h(0) > 0$. Substituting $\gamma = 1$ in (1) and setting $G(\cdot)$ and $H(\cdot)$ to be the uniform cdf's on $[0, 1]$, expressions (3) and (4) are reduced to (1) and to the density of a GTSP random variable given by Kotz and Van Dorp, 2004a.

The new pdf (4) also has the following alternative *mixture* representation:

TS-GTL distributions

$$f_Y(y|\underline{\Theta}) = p(\underline{\Theta}) \left[\frac{m}{\theta} g\left(\frac{y}{\theta}\right) \left\{ G\left(\frac{y}{\theta}\right) \right\}^{m-1} \right] + \{1 - p(\underline{\Theta})\} \left[\frac{n}{1-\theta} h\left(\frac{y-\theta}{1-\theta}\right) \left\{ 1 - H\left(\frac{y-\theta}{1-\theta}\right) \right\}^{n-1} \right], \quad (7)$$

where the mixture probability is $p(\underline{\Theta}) = \theta n / \{(1-\theta)m + \theta n\}$, as above. For an integer m , the first term in (6) is easily recognized as the *largest* order statistic distribution of a sample of size m from the rescaled distribution $G(\cdot)$ with the support $[0, \theta]$. For an integer n , the second member is seen to be the *smallest* order statistic distribution of a sample of size n from the rescaled distribution $H(\cdot)$ with support $[\theta, 1]$. This explains the designation of the parameter θ as the threshold. For non-integer $n, m > 0$, they could be interpreted as *virtual sample sizes*. (The notion of a non-integer virtual sample size corresponds to the notion of a non-integer prior sample size introduced by Ferguson in (1973).)

Jones (2004) recently investigated generalizations of the distribution of order statistics of the form

$$\{B(a, b)\}^{-1} f(x) F^{a-1}(x) \{1 - F(x)\}^{b-1}, \quad (8)$$

where as usual $B(a, b) = \Gamma(a+b)/\Gamma(a)\Gamma(b)$, $F(\cdot)$ is a particular cdf and $a, b > 0$ are not necessarily integers. Jones' (2004) generalizations are not restricted to distributions $F(\cdot)$ with a bounded support, but both members in the mixture (7) belong to his class of distributions. The generalization (7) of (1) may thus be viewed along the lines of the Jones' (2004) approach. While Jones' (2004) pdf's (8) may be viewed as generalizations of beta distributions, distributions (7) and those in Kotz and van Dorp (2004b) were generated in a sequel for *alternatives* to the beta distribution. Hence, expression (7) provides a natural link between both the approaches.

Introducing the notation y_q to indicate the q th quantile of the cdf (3) we obtain from (3) the following quantile function (or inverse cdf):

$$y_q = \begin{cases} \theta G^{-1} \left\{ \frac{q}{p(\underline{\Theta})} \right\}^{1/m}, & \text{for } 0 \leq q < p(\underline{\Theta}), \\ \theta + (1-\theta) H^{-1} \left(1 - \left\{ \frac{1-q}{1-p(\underline{\Theta})} \right\}^{1/n} \right), & \text{for } p(\underline{\Theta}) \leq q \leq 1, \end{cases} \quad (9)$$

TS-GTL distributions

where $p(\Theta)$ is given by (5). Hence, provided the cdf's $G(\cdot)$ and $H(\cdot)$ have quantile functions that can be expressed using only elementary functions, the cdf (3) will have a quantile function with the same property.

3. PDF and CDF of TS-GTL distributions

We shall now provide an example of the density construction (4) above by letting $G(\cdot)$ [$H(\cdot)$] to be a slope [reflected slope] distribution on $[0, 1]$ given by:

$$\begin{cases} G(x|\alpha) = \alpha x - (\alpha - 1)x^2, \\ H(x|\beta) = 1 - \beta(1 - x) + (\beta - 1)(1 - x)^2, \end{cases} \quad (10)$$

with the corresponding pdf's:

$$\begin{cases} g(x|\alpha) = \alpha - 2(\alpha - 1)x, \\ h(x|\beta) = \beta - 2(\beta - 1)(1 - x), \end{cases} \quad (11)$$

where $0 \leq \alpha, \beta \leq 2$. Substituting (10) and (11) into (4) we obtain the density:

$$f_Y(y|\Theta, \alpha, \beta) = \frac{mn}{(1 - \theta)m + \theta n} \times \quad (12)$$

$$\begin{cases} \left\{ \alpha - 2(\alpha - 1)\left(\frac{y}{\theta}\right) \right\} \left\{ \alpha\left(\frac{y}{\theta}\right) - (\alpha - 1)\left(\frac{y}{\theta}\right)^2 \right\}^{m-1}, & \text{for } 0 < y < \theta, \\ \left\{ \beta - 2(\beta - 1)\left(\frac{1-y}{1-\theta}\right) \right\} \left\{ \beta\left(\frac{1-y}{1-\theta}\right) - (\beta - 1)\left(\frac{1-y}{1-\theta}\right)^2 \right\}^{n-1}, & \text{for } \theta \leq y < 1, \\ 0, & \text{elsewhere,} \end{cases}$$

(where as above $\Theta = (\theta, m, n)$) with the cdf

$$F_Y(y|\Theta, \alpha, \beta) = \begin{cases} 0, & \text{for } y \leq 0, \\ p(\Theta) \left\{ \alpha\left(\frac{y}{\theta}\right) - (\alpha - 1)\left(\frac{y}{\theta}\right)^2 \right\}^m, & \text{for } 0 < y < \theta, \\ 1 - \{1 - p(\Theta)\} \left\{ \beta\left(\frac{1-y}{1-\theta}\right) - (\beta - 1)\left(\frac{1-y}{1-\theta}\right)^2 \right\}^n, & \text{for } \theta \leq y < 1, \\ 1, & \text{for } y \geq 1. \end{cases} \quad (13)$$

Setting $n = m$ and $\alpha = \beta = 2$ in (12), we arrive at the density of a Two-Sided Topp and Leone distribution. Originally Topp and Leone (1955) had introduced their distribution with specific reliability applications in mind. The Topp and Leone work was used by Nadarajah and Kotz (2003) and the TL distribution was generalized by Kotz and Van Dorp (2004b), p. 198. We shall refer to

TS-GTL distributions

the distribution defined in (12) and (13) as the *Two-Sided Generalized Topp and Leone* (TS-GTL) distribution.

Figure 1 plots some examples of the density (4), using (10), for different value settings of the parameters m, n, θ, α and β . Figures 1A, B and C plot the uniform, a triangular and GTSP distributions all with the common values $\alpha = 1, \beta = 1$. Indeed, for $\alpha = 1, \beta = 1$ distributions (10) reduce to a uniform distribution and hence the density (11) reduces to the GTSP density (1) of which the distributions depicted in Figures 1A, B and C are members. Figure 1D ($m = 1, n = 1$) plots a two-sided slope distribution (see, e.g., Kotz and Van Dorp (2004a)), which also reduces to a triangular distribution by setting $\alpha = \beta = 0$). Figure 1E plots a continuous bimodal density with $\alpha = \beta$. Amongst the shapes in Figure 1, the bimodal one in Figure 1E is a new addition to the distributional forms presented in Kotz and Van Dorp (2004b). The pdf (11) assumes this form when $\alpha, \beta \in (1, 2], n, m > 1$ with the two modes at

$$\begin{cases} x_1 = \theta \text{Min} \left[1, \frac{\alpha}{2(\alpha-1)} \left\{ 1 - \sqrt{\frac{1}{2m-1}} \right\} \right], \\ x_2 = 1 - (1 - \theta) \text{Min} \left[1, \frac{\beta}{2(\beta-1)} \left\{ 1 - \sqrt{\frac{1}{2n-1}} \right\} \right]. \end{cases} \quad (14)$$

Finally, Figure 1F plots a discontinuous density (since here $\alpha \neq \beta$) reminiscent of the uneven two-sided power (UTSP) densities. The UTSP distributions were derived in Kotz and Van Dorp (2004a) distributions of a generalized trapezoidal distribution discussed in Van Dorp and Kotz (2003b). Since the distributions in Figures 1D, E and F are not members of the family of distributions given by (1), the structure of the density (4) (combined with the cdf's (10)) constitutes a single framework for the families of distributions that were previously dispersed amongst several related but separate classes.

We obtain from (10) the following quantile functions for the cdf's $G(\cdot | \alpha), H(\cdot | \beta)$,

$$G^{-1}(q|\alpha) = \begin{cases} \frac{\alpha + \sqrt{\alpha^2 - 4(\alpha-1)q}}{2(\alpha-1)}, & \text{for } 0 \leq \alpha < 1, \\ q, & \text{for } \alpha = 1 \\ \frac{\alpha - \sqrt{\alpha^2 - 4(\alpha-1)q}}{2(\alpha-1)}, & \text{for } 1 < \alpha \leq 2, \end{cases} \quad (15)$$

TS-GTL distributions

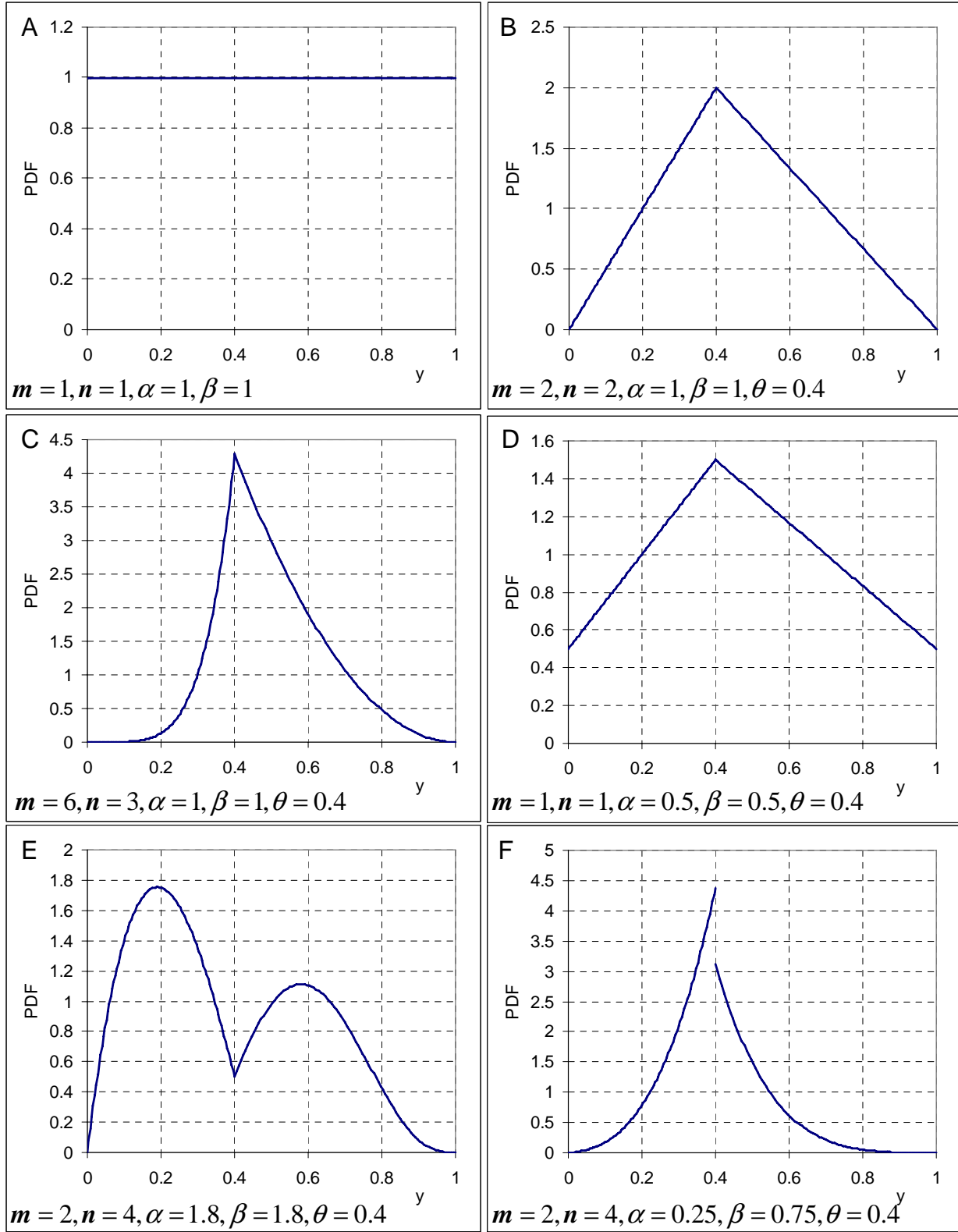


Figure 1. Examples of Two Sided - Generalized Topp and Leone (TS-GTL) distributions.

and

$$H^{-1}(q|\beta) = \begin{cases} 1 - \frac{\beta + \sqrt{\beta^2 - 4(\beta-1)(1-q)}}{2(\beta-1)}, & \text{for } 0 \leq \beta < 1, \\ q, & \text{for } \beta = 1 \\ 1 - \frac{\beta - \sqrt{\beta^2 - 4(\beta-1)(1-q)}}{2(\beta-1)}, & \text{for } 1 < \beta \leq 2. \end{cases} \quad (16)$$

where $q \in [0, 1]$ is a quantile level. Hence, the combination of expressions (15), (16) and (9) allow for a direct evaluation of the quantiles y_q of the cdf (13). Utilizing (15), (16) and (9) it follows that sampling from distributions (13) using the inverse cdf technique and a pseudo-random number generator that samples these quantile levels q (see, e.g., Banks *et al.* 2005) is straightforward and computationally efficient as well.

4. Moments

In this section we shall describe a procedure to calculate $E[Y^k|\Theta, \alpha, \beta]$ for the density (10), via the mixture structure (7). Let $X_1 [X_2]$ be a random variable with density function $mg(x) \{G(x)\}^{m-1} [nh(x) \{H(x)\}^{m-1}]$ where G and H are as defined by (10). From (7) we have :

$$E[Y^k|\Theta, \alpha, \beta] = \frac{p(\Theta)}{\theta} E[X_1^k|\alpha, m] + \frac{\{1 - p(\Theta)\}}{1 - \theta} [E[X_2^k|\beta, n] + \theta]. \quad (17)$$

Hence in the expression (17) $X_1 [X_2]$ is a [Reflected] Generalized Topp and Leone distribution with the support $[0, 1]$ (see Kotz and Van Dorp (2004a), p. 198). Alternatively, $X_1 [X_2]$ is the largest [smallest] order statistic of a random sample from a [reflected] slope distribution with parameter α [β] of virtual sample size $m [n]$ (since $m > 0 [n > 0]$ is not necessarily integer).

Jones (2004) (amongst others) notes that derivation of closed form expressions for the moments of an order statistic distribution could be somewhat complicated on a case by case basis. The moments of X_1 and X_2 are no exception. Nadarajah and Kotz (2003) in a short paper derived the moment expressions for X_1 for the case for $\alpha = 2$. These results were further generalized by Kotz

TS-GTL distributions

and Van Dorp (2004a), to derive the cumulative moments⁴ for *reflected generalized Topp and Leone* distributions for $0 \leq \beta \leq 2$ for X_2 :

$$M_k = \int_0^1 x^k (1 - H(x|\beta, n)) dx = \sum_{i=0}^k \binom{k}{i} (-1)^i \beta^n \int_0^1 x^{n+i} \left\{ 1 - \frac{(\beta-1)x}{\beta} \right\}^n dx \quad (18)$$

for $0 \leq \beta \leq 2$. The moments $\mu'_k = E[X_2^k|\beta, n]$ are connected with the cumulative moments M_k , $k = 1, \dots, 4$, via the well known relationship :

$$\mu'_k = kM_{k-1}, k = 1, 2, 3, \dots \quad (19)$$

(see, e.g., Stuart and Ord (1994)). For $\beta \in (1, 2]$, the cumulative moments can further be expressed utilizing the incomplete beta function $B(x | a, b) = \mathbb{B}^{-1}(a, b) \int_0^x p^{a-1} (1-p)^{b-1} dp$:

$$M_k = \sum_{i=0}^k \binom{k}{i} (-1)^i \beta^n \left\{ \frac{\beta}{\beta-1} \right\}^{n+i+1} \frac{B\left(\frac{\beta-1}{\beta} | n+i+1, n+1\right)}{\mathbb{B}^{-1}(n+i+1, n+1)} \quad (20)$$

where as above $\mathbb{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. Explicit recursive relations for M_k , $k = 1, 2, 3$, can be easily derived from (20). Numerical routines for evaluating the incomplete beta function are well known from Pearson's investigations in early 20th century and are provided in standard PC software such as e.g. Microsoft Excel. Unfortunately, for $\beta \in (0, 1)$ the expression (14) cannot be further simplified and one has to resort to numerical integration. Noting that the pdf of X_1 is a generalized Topp and Leone distribution and that of X_2 is a reflected one, we have in turn the following relationship:

$$E[X_1^k|\alpha, m] = E[(1 - X_2)^k|\beta, n] \Big|_{\beta=\alpha, n=m} = \sum_{i=0}^k \binom{k}{i} (-1)^i E[X_2^i|\alpha, m]. \quad (21)$$

Summarizing, the mean, variance, skewness and kurtosis can straightforwardly be evaluated by means of an algorithm that utilizes expressions (19), (20) and (21) for $\alpha, \beta \in (1, 2]$. For example,

⁴Kotz and Van Dorp (2004a) considered cumulative moments $M_k = \int_0^1 x^k (1 - F(x)) dx$ since they were studying the reflected generalized TL distributions (unlike Nadarajah and Kotz (2003) who were dealing with the moments of TL distributions).

TS-GTL distributions

applying the procedure above for $\alpha, \beta \in (1, 2]$ and $k = 1$ we arrive at the following expression for the mean of Y with pdf (12):

$$E[Y|\Theta, \alpha, \beta] = \frac{p(\Theta)}{\theta} \left[1 - \alpha^m \left\{ \frac{\alpha}{\alpha - 1} \right\}^{m+1} \frac{B(\frac{\alpha-1}{\alpha} | m+1, m+1)}{\mathbb{B}^{-1}(m+1, m+1)} \right] + \frac{1 - p(\Theta)}{1 - \theta} \left[\beta^n \left\{ \frac{\beta}{\beta - 1} \right\}^{n+1} \frac{B(\frac{\beta-1}{\beta} | n+1, n+1)}{\mathbb{B}^{-1}(n+1, n+1)} + \theta \right]. \quad (22)$$

For $\alpha = \beta = \{0, 1\}$, the moments of $E[X_1^k|\alpha, m]$ and $E[X_2^k|\beta, n]$ are reduced to those of a power and reflected power distributions which are available in a closed form (see, e.g., Kotz and Van Dorp (2004a)). For $\alpha, \beta \in (0, 1)$ evaluations of cumulative moments $E[X_1^k|\alpha, m]$ and $E[X_2^k|\beta, n]$ require numerical integrations techniques employing (18).

5. Maximum Likelihood Procedure

For a random sample $\underline{X} = (X_1, \dots, X_s)$ of size s from the distribution (12), the loglikelihood function is, by definition,

$$\begin{aligned} \text{Log}\{L(\underline{X}, \Theta, \alpha, \beta)\} &= s \text{Log}\left\{ \frac{mn}{(1-\theta)m + \theta n} \right\} + \\ &\sum_{i=1}^r \text{Log}\left\{ g\left(\frac{X_{(i)}}{\theta} | \alpha\right) \right\} + (m-1) \sum_{i=1}^r \text{Log}\left\{ G\left(\frac{X_{(i)}}{\theta} | \alpha\right) \right\} + \\ &\sum_{i=r+1}^s \text{Log}\left\{ h\left(\frac{X_{(i)} - \theta}{1-\theta} | \beta\right) \right\} + (n-1) \sum_{i=r+1}^s \text{Log}\left\{ 1 - H\left(\frac{X_{(i)} - \theta}{1-\theta} | \beta\right) \right\}, \end{aligned} \quad (23)$$

where $g(\cdot | \alpha)$, $G(\cdot | \alpha)$, $h(\cdot | \alpha)$, $H(\cdot | \alpha)$ are defined by (10) and (11), $X_{(1)} < X_{(2)} < \dots < X_{(s)}$ are the order statistics of \underline{X} , and r is a positive integer such that

$$X_{(r)} \leq \theta < X_{(r+1)}. \quad (24)$$

By convention $X_{(0)} = -\infty$, $X_{(s+1)} = +\infty$. We propose the following algorithm to maximize the loglikelihood $\text{Log}\{L(\underline{X}, \Theta, \alpha, \beta)\}$ (23) and to determine the ML estimates of the parameters m, n, α, β and θ using a feasible starting point $m^*, n^*, \alpha^*, \beta^*$ and θ^* :

The k th iteration:

Step 0: Set $k = 1$, $\alpha_1 = \alpha^*$, $\beta_1 = \beta^*$, $m_1 = m^*$, $n_1 = n^*$, $\theta = \theta^*$.

TS-GTL distributions

Step 1: Determine m_{k+1} by maximizing $\text{Log}\{L(\underline{X}|m, n_k, \alpha_k, \beta_k, \theta_k)\}$ over m .

Step 2: Determine n_{k+1} by maximizing $\text{Log}\{L(\underline{X}|m_{k+1}, n, \alpha_k, \beta_k, \theta_k)\}$ over n .

Step 3: Determine α_{k+1} by maximizing $\text{Log}\{L(\underline{X}|m_{k+1}, n_{k+1}, \alpha, \beta_k, \theta_k)\}$ over α .

Step 4: Determine β_{k+1} by maximizing $\text{Log}\{L(\underline{X}|m_{k+1}, n_{k+1}, \alpha_{k+1}, \beta, \theta_k)\}$ over β .

Step 5: Determine θ_{k+1} by maximizing $\text{Log}\{L(\underline{X}|m_{k+1}, n_{k+1}, \alpha_{k+1}, \beta_{k+1}, \theta)\}$ over θ .

Step 6: If $|\text{Log}\{L(\underline{X}|m_{k+1}, n_{k+1}, \alpha_{k+1}, \beta_{k+1}, \theta_{k+1})\} - \text{Log}\{L(\underline{X}|m_k, n_k, \alpha_k, \beta_k, \theta_k)\}| < \epsilon$

STOP

Else $k = k + 1$ and Goto Step 1.

To obtain an initial starting solution $m^*, n^*, \alpha^*, \beta^*$ and θ^* in Step 0 one could, for example, select $m^*, n^*, \alpha^*, \beta^*$ and θ^* visually to match a plot of a TS-GTL pdf to that of an empirical pdf or more directly use least squares estimates for $m^*, n^*, \alpha^*, \beta^*$ and θ^* . As usual, ϵ in Step 6 may be chosen arbitrarily small.

Setting the partial derivatives of (23) with respect to left branch power parameter m equal to 0, we obtain the following MLE for m_{k+1} at Step 1:

$$\widehat{m}(\theta, n) = \frac{1}{2} \frac{\theta n}{1 - \theta} \left\{ -1 + \sqrt{1 + \mathcal{K}(\theta, n) \frac{4s(1 - \theta)}{n\theta}} \right\} > 0, \quad (25)$$

where

$$\mathcal{K}(\theta, n) = \left[- \sum_{i=1}^r \text{Log}\left\{G\left(\frac{X_{(i)}}{\theta} \mid \alpha\right)\right\} \right]^{-1} > 0. \quad (26)$$

Hence, the MLE $\widehat{m}(\theta, n)$ maximizes the log-likelihood profile of (23) as a function of m , keeping the remainder of the parameters in (23) fixed. Analogously, we obtain for the right branch power parameter n_{k+1} at Step 2 of the algorithm above:

$$\widehat{n}(\theta, m) = \frac{1}{2} \frac{(1 - \theta)m}{\theta} \left\{ -1 + \sqrt{1 + \mathcal{Z}(\theta, m) \frac{4s\theta}{m(1 - \theta)}} \right\} > 0, \quad (27)$$

where

TS-GTL distributions

$$\mathcal{Z}(\theta, n) = \left[- \sum_{i=r+1}^s \text{Log} \left\{ 1 - H \left(\frac{X_{(i)} - \theta}{1 - \theta} \mid \beta \right) \right\} \right]^{-1} > 0. \quad (28)$$

Both expressions (25) and (27) depend on the form of the cdf's $G(\cdot)$ and $H(\cdot)$ only via the expressions (26) and (28) and thus also apply to the general structure of the pdf (4).

Determination of the ML estimates of the log-likelihood profile of (23) as a function of α , β or θ turns out to be more challenging. Setting the partial derivative of (23) with respect to α equal to zero requires solving the following equation for $\alpha \in [0, 2]$:

$$\sum_{i=1}^r [\gamma_i - \zeta_i(\alpha)] \xi_i(\alpha) = 0, \quad (29)$$

where

$$\begin{aligned} \gamma_i &= m(\theta - X_{(i)}) > 0, \\ \xi_i(\alpha) &= [\alpha(\theta - X_{(i)}) + X_{(i)}]^{-1} > 0, \\ \zeta_i(\alpha) &= \theta X_{(i)} / [\alpha\theta - 2(\alpha - 1)X_{(i)}] > 0, \end{aligned} \quad (30)$$

for $i = 1, \dots, r$ (where as above r is a positive integer defined by (24)). A similar equation may be derived for the log-likelihood profile of (23) as a function of β . Figure 2A plots a log-likelihood profile of (23) as a function of α , which indicates that multiple solutions to equation (29) may possibly exist, whereas the global optimum over $\alpha \in [0, 2]$ is actually attained at $\alpha = 0$. Figure 2B plots a log-likelihood profile of (23) as a function of θ , which shows (i) a global optimum at the lower bound of the range $\theta \in [0, 1]$, (ii) a discontinuous behavior of the log-likelihood as a function of θ over $[0, 1]$, but continuous over each interval $[X_{(i)}, X_{(i+1)}]$, $i = 0, \dots, s$ (where s is the sample size) and (iii) the existence of stationary points within the interval $[X_{(i)}, X_{(i+1)}]$.

Given the structure of the log-likelihood profiles as a function of α , β or θ , it would seem reasonable to take advantage of the boundedness of $\alpha, \beta \in [0, 2]$ and $\theta \in [0, 1]$ and globally optimize over these intervals by discretizing at a desirable level of accuracy δ and to evaluate the log-likelihood at all discretized points. In the case of θ , and given the behavior of (23) in e.g. Figure 2B at the order statistics $X_{(i)}$, we recommend to include in this optimization procedure also the

TS-GTL distributions

evaluation of the log-likelihood (23) at the order statistics $X_{(i)}$. (Recall that in the special case that (23) reduces to a log-likelihood of a TSP distribution, Kotz and Van Dorp (2004a, p. 80) have shown that the MLE $\hat{\theta}$ is attained at one of the order statistics $X_{(i)}$, $i = 1, \dots, s$, provided $n > 1$.) While the above optimization procedure for the log-likelihood profiles as a function of α , β or θ is not particularly elegant, it certainly is practical given current available computational facilities. An EXCEL spreadsheet program with an implementation of the algorithm above is available from the authors upon request.

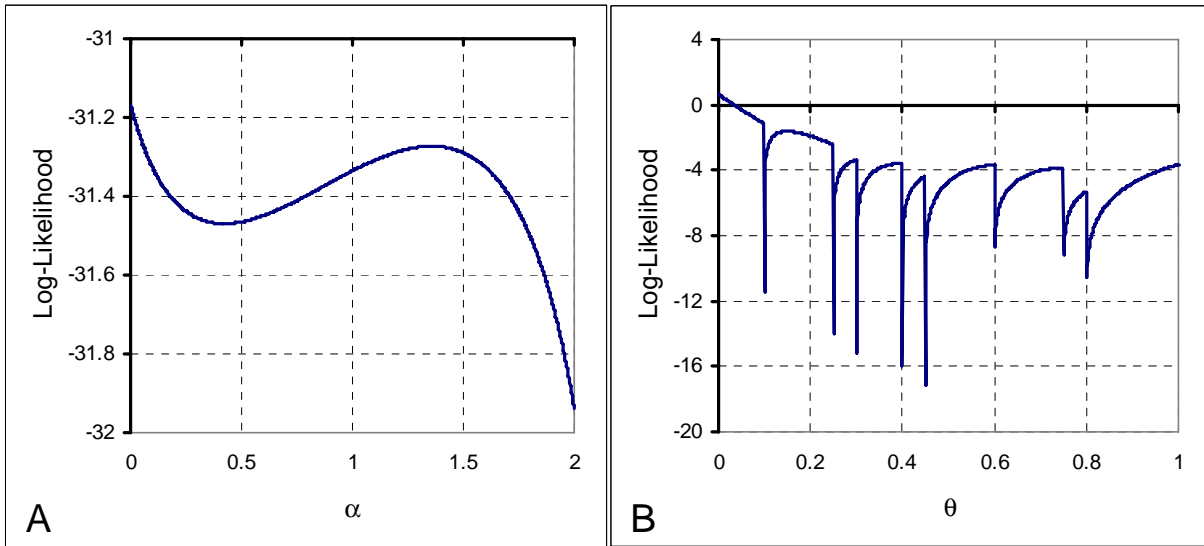


Figure 2. Examples of log-likelihood profiles as a function of α (Fig. A) and as a function of θ (Fig. B).

The algorithm above can be easily modified to a ML algorithm for sub-classes in the TS-GTL family. For example, omitting Steps 1, 2 and 4 and setting $m = n = 1$, $\beta = \alpha$ results in a ML algorithm for the TS-Slope distributions. Setting $\alpha = 1$, $\beta = 1$ and removing Steps 3 and 4, we obtain an ML algorithm for GTSP distributions. Next, by setting $m = n$ and removing Step 2 ($m = n = 2$ and removing Steps 1 and 2) it reduces to an algorithm for the TSP (triangular) distribution. Finally, eliminating just Step 4, while guaranteeing $\beta = \alpha$ leads to an ML algorithm for continuous TS-GTL distributions. Various versions of the ML algorithm will be used in the example presented in the next section.

6. An Illustrative Example

The classical data in this example reaches the high points of our gigantic Universe and involves the $(V - I)$ indices of 80 globular clusters in the Galaxy M87. Galaxy M87, also called Virgo A, was discovered by Charles Messier, a French astronomer, in 1781 (see, e.g., Philbert (2000)). Note that we are dealing here with data listed in Harris (2003) that originated outside the Galaxy containing our solar system. Davis and Brodie (2006) explain: "Globular clusters are nearly spherical groups of about 10,000 to 1 million stars. The color of a globular cluster gives clues about the cluster's composition (what kinds of elements and stars are in the cluster) and the cluster's age. ... V and I are different filters through which we can look at objects in the sky. Looking through a V filter is like looking through a yellow pair of glasses and looking through an I filter is like looking through infrared glasses (our eyes can't see infrared, but telescopes can)." Hence, a $(V - I)$ index is a color measurement index. The smallest (largest) $(V - I)$ index in the data set equals 0.767 (1.293), hence we translate the data $\frac{1}{2}$ unit to the left resulting in order statistics $X_{(1)} = 0.267$, $X_{(80)} = 0.793$ to obtain a data set within $[0, 1]$. (Recall that TS-GTL distributions have support $[0, 1]$). We next, fit TS-GTL distribution to the translated data set using the ML algorithm described in the previous section. To obtain distributions for the original data set one simply shifts the distribution $\frac{1}{2}$ unit to the right.

Table 1 provides the ML estimates for the various distributions that were fitted, all members of the TS-GTL family of distributions. Those entries in Table 1 indicated with * were fixed in the ML algorithm to ensure that a member within a particular sub-class of the TS-GTL was fitted. For example, in the first row only θ is a free variable, resulting in the ML fit of the triangular distribution. Note that the ML fit of the TSS distribution actually coincides with the triangular one in the first row. Indeed for both parameter settings, $m = n = 2$, $\alpha = \beta = 1$ and $m = n = 1$, $\alpha = \beta = 0$, the pdf (12) reduces to a triangular pdf.

Figure 3 provides QQ-plots for the ML fitted triangular, GTSP, TS-GTL (continuous), $\alpha = \beta$ in (12), and TS-GTL (discontinuous), $\alpha \neq \beta$ in (12), distributions in Table 1. The distributions in the first and second row of Table 1 are identical, whereas the QQ-plot for the TSP and the GTSP is

TS-GTL distributions

quite similar with only a slight improvement for the GTSP distribution over the TSP distribution. From the QQ-plots we observe a better fit of the GTSP distribution over the triangular one, but neither seem to perform particularly well. The QQ-plots of the TS-GTL (continuous) and TS-GTL (discontinuous), however, align quite well with the $x = y$ line. It thus appears from the QQ-plots that both the TS-GTL (continuous) and TS-GTL (discontinuous) are "reasonable" eyeball fits for the data under consideration.

Table 1. ML parameter estimates of fitted distributions to the Galaxy M87 (Virgo A) data. Those parameters indicated with * are fixed or of the same value as another parameter

(e.g. $n = m = 2$, $\alpha = \beta = 1$ in the first row).

	m	n	θ	α	β
Triangular	2*	2*	0.626	1*	1*
TS - Slope (TSS)	1*	1*	0.626	0	0*
TS - Power (TSP)	3.496	3.496*	0.626	1*	1*
Gen. TS - Power (GTSP)	3.288	3.987	0.635	1*	1*
TS - GTL (Cont.)	12.843	8.978	0.549	1.948	1.948*
TS - GTL (Discont.)	11.724	8.249	0.582	2.0	1.758

Figure 4 displays an empirical kernel density of the translated Galaxy M87 ($V - I$) index color data. The empirical density was generated using the Bartlett-Epanechnikov kernel (e.g., Izenman, 1991) combined with the over-smoothed bandwidth (e.g., Sheather, 2004). We observe that our data set seems to be bimodal which may explain why the QQ-plots of the triangular and GTSP distributions in Figure 3 are inappropriate. From Figure 4C we observe that the TS-GTL (continuous) distribution captures both the mode locations and the anti-mode location of the empirical density. Note that it overestimates (underestimates) the peak of the right (left) mode, as if having difficulties to capture the width (narrowness) of the right (left) part of the empirical density. Similar observations can be drawn for the TS-GTL (discontinuous) from Figure 4D, but adding a degree of freedom (allowing for a discontinuity at the anti-mode) seems to result in a less favorable fit than the one in Figure 4C.

TS-GTL distributions

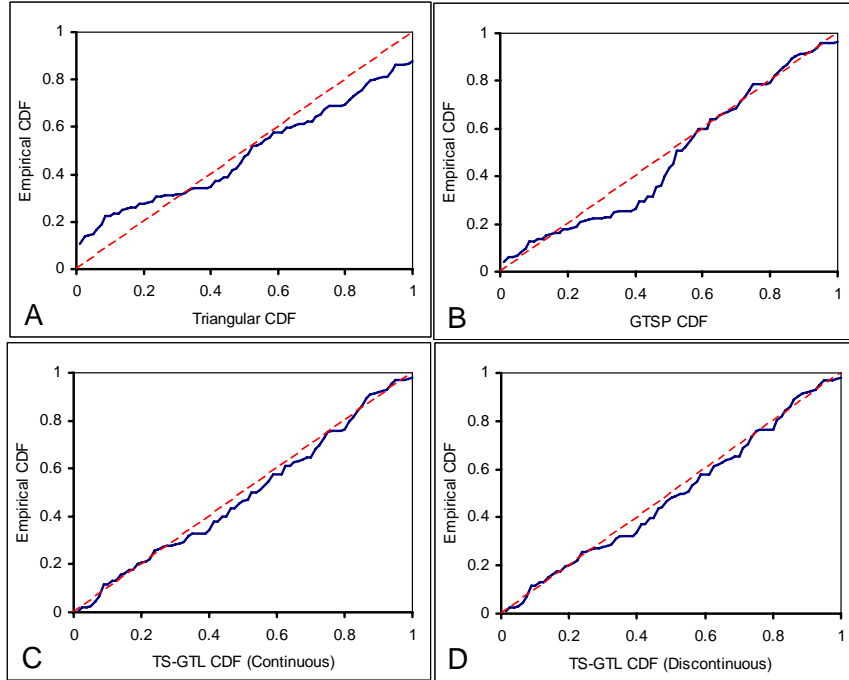


Figure 3. QQ-plots of four ML fitted distributions in Table 1 to Galaxy M87 (Virgo A) data:
 A: triangular; B: GTSP; C: TS-GTL (Continuous); D: TS-GTL (Discontinuous).

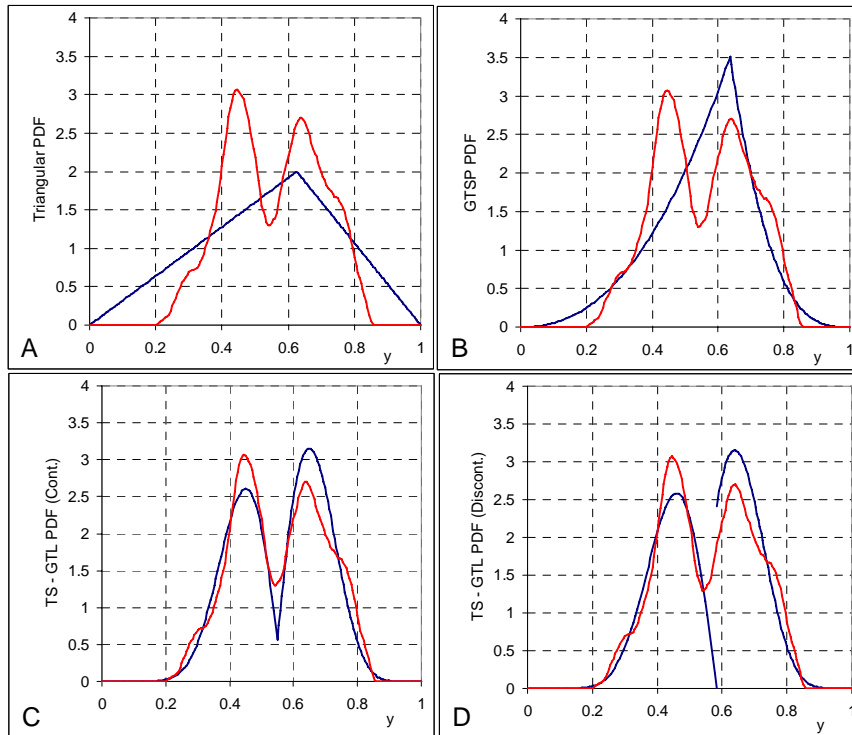


Figure 4. Empirical PDF and PDF's of four ML fitted distributions in Table 1 to Galaxy M87 (Virgo A) data: A: triangular; B: GTSP; C: TS-GTL (Continuous); D: TS-GTL (Discontinuous).

TS-GTL distributions

To obtain a deeper appreciation of the data, we have decided also to fit a Gaussian mixture model with two components

$$\pi\phi(x|\mu_1, \sigma_1) + (1 - \pi)\phi(x|\mu_2, \sigma_2) \quad (31)$$

using an Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin (1993)), where $\phi(\cdot|\mu_i, \sigma_i)$ are Gaussian pdf's and π and $1 - \pi$ are the mixture weights. The ECM algorithm also provides a maximum likelihood estimation. Figure 5 plots the ML fitted Gaussian mixture pdf and the QQ plot with the MLE's:

$$\hat{\pi} = 0.500, \hat{\mu}_1 = 0.437, \hat{\sigma}_1 = 0.071, \hat{\mu}_2 = 0.672, \hat{\sigma}_2 = 0.070. \quad (32)$$

Comparing Figures 5A, 4C and 4D we observe a similar behavior, namely that the Gaussian mixture fit underestimates (overestimates) the pdf value at the first (second) mode. On the other hand, the Gaussian mixture captures better the value of the empirical pdf at the anti-mode than the fitted pdf's in Figures 4C and 4D. This perhaps partially explains the situation that the QQ-plot 5B seems to be more accurate at the center than the QQ-plots in Figures 3C and 3D. (It is of course a personal judgment in each case whether the center or tails are more important.)

The results in the QQ-plots in Figures 3 and 5 and our conjectures from the comparison of Figures 4C, 4D and 5 are now tested by a formal fit-analysis presented in Table 2. The Chi-square statistic in Table 2 is calculated utilizing 10 bins as suggested by Banks et al. (2001). The boundaries of the bins are selected such that the number of observations $O_i, i = 1, \dots, 15$, in each bin i is equal to 7, 8 or 9, totaling 80 data points. Such a boundary selection procedure partitions the support of the range of observed data along the lines of the "equal probability method of constructing classes" (e.g., Stuart and Ord, 1994). The TS-GTL (continuous) distribution results in the largest p -value(0.32) of the Chi-square test of the goodness of fit. The TS-GTL (discontinuous) provides the second largest (0.16) and the Gaussian mixture pdf gives the third largest (0.13, only insignificantly smaller than the TS-GTL(discontinuous) case). The other distributions yield negligible p -values (way below 0.01). Table 2 also contains the BIC-criterion [AIC-criterion] indicating a better [worse] fit for the TS-GTL (continuous) as compared to the TS-GTL(discontinuous) distribution.

TS-GTL distributions

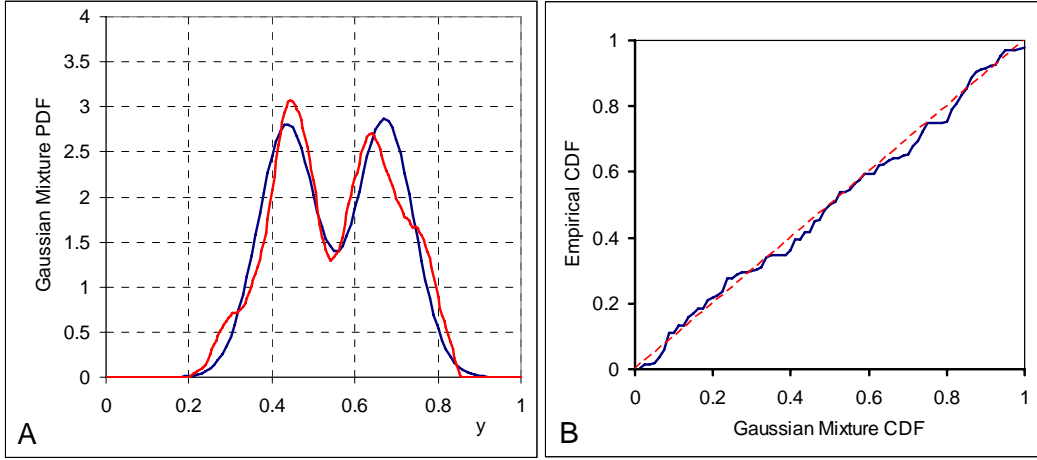


Figure 5. A: Empirical PDF and ML fitted Gaussian mixture PDF to Galaxy M87 (Virgo A) data
 B: QQ-plots of ML fitted Gaussian mixture PDF with MLE's given in (32).

Table 2. Fit analysis of ML fitted distributions to Galaxy M87 (Virgo A) data.

Bin	LB_i	UB_i	O_i	triangular	TSP	GTSP	TS - GTL (Cont.)	TS - GTL (Dis.)	Gaussian mixture
				$(O_i - E_i)^2 / E_i$	$(O_i - E_i)^2 / E_i$	$(O_i - E_i)^2 / E_i$	$(O_i - E_i)^2 / E_i$	$(O_i - E_i)^2 / E_i$	$(O_i - E_i)^2 / E_i$
1	0.000	0.383	8	6.13	0.10	0.47	0.14	0.13	0.10
2	0.383	0.425	8	3.10	4.19	3.53	0.05	0.12	0.02
3	0.425	0.455	8	6.35	5.87	5.32	0.55	0.66	0.25
4	0.455	0.478	8	10.09	7.82	7.37	2.36	2.28	2.16
5	0.478	0.556	8	0.52	2.29	2.32	0.33	1.19	0.92
6	0.556	0.613	9	0.03	1.47	1.32	0.01	0.28	0.27
7	0.613	0.638	7	2.36	0.01	0.01	0.18	0.10	1.13
8	0.638	0.674	8	1.38	0.00	0.05	0.10	0.08	0.00
9	0.674	0.740	8	0.01	0.45	0.37	1.75	1.46	1.94
10	0.740	1.000	8	2.91	0.02	0.22	0.38	0.33	0.24
Chi-square statistic				32.87	22.23	20.99	5.86	6.64	7.02
Parameters				1	2	3	4	5	5
Degrees of freedom				8	7	6	5	4	4
p-value				6.5E-05	2.3E-03	1.8E-03	0.32	0.16	0.13
AIC - criterion				-63.14	-82.04	-80.91	-98.00	-98.59	-94.38
BIC - criterion				-60.76	-77.28	-73.77	-88.48	-86.68	-82.47

The Gaussian mixture pdf is slightly outperformed by the TS-GTL (continuous) and the TS-GTL(discontinuous) pdf's in both the AIC and BIC-criteria. Finally, we have a p -value of 0.107 for the likelihood-ratio test with 1 d.f. for the nested TS-GTL (discontinuous) distribution within the

TS-GTL distributions

TS-GTL (discontinuous) distribution. Thus, the overall fit analysis convincingly suggests that the TS-GTL (continuous) results in the best fit amongst the six distributions tested in Table 2.

It may be interesting to note that the threshold parameter θ of the pdf (7) allows for a direct 0-1 classification of each data point to either one of its mixing subpopulations, while in the case of the Gaussian mixtures (31), the above mentioned ECM algorithm yields estimated probabilities of a data point belonging to each mixing subpopulation, allowing for an indirect 0-1 classification of data points.

References

The authors are thankful to the referee and the editor of the Journal of Applied Statistics whose valuable comments improved the contents and presentation of an earlier version.

References

- Banks, J., Carson, J.S., Nelson, B.L. and Nicol, D.M. (2005). *Discrete-Event System Simulation* (4th ed.), Prentice-Hall, Upper Saddle River, NJ.
- Davis, E., and Brodie, J. (2006). *The Milky Way and Beyond: Globular Clusters*, <http://www.sciencebuddies.org> (ed. A. Olson), Science Buddies.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1: 209-230.
- Harris, W.E. (2003). *Catalog Parameters for Milky Way Globular Clusters: The Database*, McMaster University [accesses April 21, 2006].
- Izenman, A.J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86 (413):205 – 224.
- Jones, M.C. (2004). Families of distributions arising from distributions of order statistics. *Test*, 13 (1): 1-43.
- Kotz, S. and Van Dorp, J.R. (2004a). Uneven two-sided power distributions: with applications in econometric models". *Statistical Methods and Applications*, 13: 285-313.

TS-GTL distributions

- Kotz, S. and Van Dorp, J.R. (2004b). *Beyond Beta, Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific Press, Singapore.
- Meng, X.L., Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80: 267-278.
- Nadarajah, S., and Kotz, S. (2003), Moments of some J-shaped distributions. *Journal of Applied Statistics*, 30 (3): 311-317.
- Pairman, E. and Pearson, K. (1919). On corrections for the moment-coefficients of limited range frequency distributions when there are finite or infinite ordinates and any slopes at the terminals of the range. *Biometrika*, 12 (3/4): 231-258.
- Philbert, J. (2000). *Charles Messier, Le Furet des Cometes*. Edition Pierron, Sarreguemines, France.
- Sheather, S.J. (2004). Density estimation. *Statistical Science*, 19 (4): 588 – 597.
- Stuart, A. and Ord, J.K. (1994), *Kendall's Advanced Theory of Statistics* (Vol. 1., Distribution Theory). Wiley, New York.
- Topp, C.W., and Leone, F.C. (1955). A family of J-shaped frequency functions. *Journal of the American Statistical Association*, 50 (269): 209-219.
- Van Dorp, J.R., and Kotz, S. (2002). The standard two sided power distribution and its properties: with applications in financial engineering, *The American Statistician*, 56 (2): pp. 90-99.
- Van Dorp, J.R. and Kotz, S. (2003a). Generalizations of Two Sided Power Distributions and their Convolution, *Communications in Statistics: Theory and Methods*, 32 (9): pp. 1703 - 1723.
- Van Dorp, J.R. and Kotz, S. (2003b). Generalized trapezoidal distributions. *Metrika*, 58 (1): 85-97.